DEFENSE TECHNICAL INFORMATION CENTER

Information for the Defense Community

DTIC® has determined on 11/14/2018 that this Technical Document has the Distribution Statement checked below. The current distribution for this document can be found in the DTIC® Technical Report Database.

☒ **DISTRIBUTION STATEMENT A.** Approved for public release; distribution is unlimited.

☐ **© COPYRIGHTED**; U.S. Government or Federal Rights License. All other rights and uses except those permitted by copyright law are reserved by the copyright owner.

☐ **DISTRIBUTION STATEMENT B.** Distribution authorized to U.S. Government agencies only (fill in reason) (date of determination). Other requests for this document shall be referred to (insert controlling DoD office)

☐ **DISTRIBUTION STATEMENT C.** Distribution authorized to U.S. Government Agencies and their contractors (fill in reason) (date of determination). Other requests for this document shall be referred to (insert controlling DoD office)

☐ **DISTRIBUTION STATEMENT D.** Distribution authorized to the Department of Defense and U.S. DoD contractors only (fill in reason) (date of determination). Other requests shall be referred to (insert controlling DoD office).

☐ **DISTRIBUTION STATEMENT E.** Distribution authorized to DoD Components only (fill in reason) (date of determination). Other requests shall be referred to (insert controlling DoD office).

☐ **DISTRIBUTION STATEMENT F.** Further dissemination only as directed by (inserting controlling DoD office) (date of determination) or higher DoD authority.

Distribution Statement F is also used when a document does not contain a distribution statement and no distribution statement can be determined.

☐ **DISTRIBUTION STATEMENT X.** Distribution authorized to U.S. Government Agencies and private individuals or enterprises eligible to obtain export-controlled technical data in accordance with DoDD 5230.25; (date of determination). DoD Controlling Office is (insert controlling DoD office).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
08/31/2010	Final report	1/8/07 to 5/31/10

4. TITLE AND SUBTITLE
Combinatorial Statistics on Trees and Networks

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-07-1-0506

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Mossel, Elchanan

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Regents of the University of California
c/o Sponsored Projects Office
2150 Shattuck Avenue, Suite 313
University of California, Berkeley, CA 94704-5940

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*

DOD ONR: Combinatorial Statistics on Trees and Networks (N0014-07-1-05-06) Final Report, Sep 2010

Elchanan Mossel*

September 29, 2010

1 Summary

We have made substantial progress in a number of areas covered by the proposal. We have shown how to efficiently reconstruct order-based distributions from noisy comparisons and noisy orders thus providing important examples where the generating model is not a tree but can still be recovered efficiently. We have made some important discoveries on reconstruction of Markov random fields, essentially showing that Markov Random Fields of bounded degrees are reconstructible if the data at all nodes is observable and that the problem is computationally intractable if some nodes are hidden.

For trees, we have studied a number of questions concerning aggregations of information from different sources, including mixtures of tree distributions and an analysis of gene trees and the corresponding population tree.

We have continued studying optimization over networks, in particular, social networks. Our results concern optimization of spread processes, the structure of Nash Equilibrium in random games, efficient greedy learning in networks from Gaussian signals and the construction of optimal local compression graphs.

Finally, we have made considerable progress in studying local algorithms for calculating marginal distributions over networks. Most importantly, we have shown that MCMC methods are applicable to graphs that are sparse on average. Indeed for the famous Ising model we have found the first tight conditions which ensure convergence.

*U.C. Berkeley. E-mail: mossel@stat.berkeley.edu

2 Progress in recovering graph-based distributions from samples

In a joint work with Mark Braverman we have shown how a structure of a linear network, i.e. an order, may be recovered efficiently from noisy comparison between elements. An extended abstract announcing the results is to appear in proceedings of SODA 2008.

Noisy Sorting without resampling [3, 4]. We study problems of inferring order given noisy information. In these problems there is an unknown order (permutation) π on n elements denoted by $1, \ldots, n$. We assume that information is generated in a way correlated with π. The goal is to find a maximum likelihood π^* given the information observed. We consider two different types of observations: noisy comparisons and noisy orders.

- *Noisy Orders (also called the Mallow's model).* Given the original permutation π, the probability of a permutation σ being generated is proportional to $e^{-\beta d_K(\sigma,\pi)}$. In other words, the probability is inverse exponential in the Kemeny distance of π from σ, which is the number of pairs ordered in π differently from σ:

  $$d_K(\pi,\sigma) = \#\{(i,j) : \pi(i) < \pi(j) \text{ and } \sigma(i) > \sigma(j)\}.$$

  We assume that we are given $\sigma_1, \ldots, \sigma_r$ that are generated independently conditioned on π.

- *Noisy Comparisons.* The input is the status of $\binom{n}{2}$ queries of the form $q(i,j)$, for $i < j$, where $q(i,j) = +(-)$ with probability $1/2 + \langle$ if $\pi(i) > \pi(j)(\pi(i) < \pi(j))$ for all pairs $i \neq j$, where $\langle > 0$ is a constant. It is assumed that the errors are independent. More generally, the input may be any collection of independent biased signals on the order relationship between pairs of elements.

In our results we present polynomial time algorithms for solving both problems with high probability. For noisy orders the running time of the algorithm is $n^{1+O((\beta r)^{-1})}$, and for noisy comparisons the algorithm runs in time $n^{O(\langle^{-3-\epsilon})}$. Both algorithms have $O(n \log n)$ query complexity (with the constant depending on $\langle, \beta$ and r).

As part of our proof we show that for both models the maximum likelihood solution π^* is close to the original permutation π. More formally, with high probability it holds that

$$\sum_i |\pi(i) - \pi^*(i)| = \Theta(n), \quad \max_i |\pi(i) - \pi^*(i)| = \Theta(\log n).$$

Our results are of interest in applications to ranking, such as ranking in sports, or ranking of search items based on comparisons by experts.

In a joint work with Matsen and Steel we analyze when can a mixture of tree-based distributions can be recovered from samples of the mixture.

Mixed-up trees: the structure of phylogenetic mixtures [10]. In this paper we apply new geometric and combinatorial approaches to the study of phylogenetic mixtures. The focus of the geometric approach is to describe the geometry of phylogenetic mixtures for the two state random cluster model, which is a generalization of the two state symmetric (CFN) model. In particular, we show that the set of mixture distributions forms a convex

polytope and we calculate its dimension; corollaries include a simple criterion for when a mixture of branch lengths on the star tree can mimic the probability distribution on splits of a resolved quartet tree. Furthermore, by computing volumes of polytopes we can clarify how "common" non-identifiable mixtures are under the CFN model. We also present a new combinatorial result which extends any identifiability result for a specific pair of trees of size six to arbitrary pairs of trees. Next we present a positive result showing identifiability of rates-across-sites models. Finally, we answer a question raised in a previous paper concerning mixed branch repulsion on trees larger than quartet trees under the CFN model.

In another phylogenetic related project with Sebastien Roch, we show how to integrate information from different gene trees in order to recover the specie tree.

Incomplete Lineage Sorting: Consistent Phylogeny Estimation From Multiple Loci [13]. We introduce a simple algorithm for reconstructing phylogenies from multiple gene trees in the presence of incomplete lineage sorting, that is, when the topology of the gene trees may differ from that of the species tree. We show that our technique is statistically consistent under standard stochastic assumptions, that is, it returns the correct tree given sufficiently many unlinked loci. We also show that it can tolerate moderate estimation errors.

We have shown how Markov Random Fields can be reconstructed from observations for bounded degree models.

Reconstruction of Markov Random Fields from Samples: Some Observations and Algorithms [5] Markov random fields are used to model high dimensional distributions in a number of applied areas. Much recent interest has been devoted to the reconstruction of the dependency structure from independent samples from the Markov random fields. We analyze a simple algorithm for reconstructing the underlying graph defining a Markov random field on n nodes and maximum degree d given observations. We show that under mild non-degeneracy conditions it reconstructs the generating graph with high probability using $\Theta(d \log n)$ samples which is optimal up to a multiplicative constant. Our results seem to be the first results for general models that guarantee that *the* generating model is reconstructed. Furthermore, we provide explicit $O(n^{d+2} \log n)$ running time bound. In cases where the measure on the graph has correlation decay, the running time is $O(n^2 \log n)$ for all fixed d. We also discuss the effect of observing noisy samples and show that as long as the noise level is low, our algorithm is effective. On the other hand, we construct an example where large noise implies non-identifiability even for generic noise and interactions. Finally, we briefly show that in some cases, models with hidden nodes can also be recovered.

In contrast when some of the nodes are hidden the problem is computationally hard.

The complexity of distinguishing Markov Random Fields [2] Markov random fields are often used to model high dimensional distributions in a number of applied areas. They provide a compact representation of the distribution that is polynomial in the size of the dependency graph and at most exponential in the size of the maximal clique in the graph.

A number of recent papers have studied the problem of reconstructing a dependency graph of bounded degree from independent samples from the Markov random field. These

results require observing samples of the distribution at all nodes of the graph. It was heuristically recognized that the problem of reconstructing the model where there are hidden variables (some of the variables are not observed) is much harder.

Here we prove that the problem of reconstructing bounded-degree models with hidden nodes is hard. More specifically we show that

- Unless $NP = RP$ it is impossible to decide in random polynomial time if two models generate distributions whose total variation distance is at most $1/3$ or at least $2/3$.
- Given two generating models whose total variation is promised to be at least $1/3$, and an oracle access to a sampler that generates independent samples from one of the two models, it is impossible to decide in randomized polynomial time which of the two samples is consistent with the model unless $NP = RP$.

These two results are tight as the two problems are solvable in polynomial time if $NP = RP$. For the second problem of deciding which of two models generates a distribution, we further consider the case where the two distributions are efficiently samplable and show that in this case if it has an efficient polynomial-time algorithm then every problem that has a zero-knowledge proofs is decidable in random polynomial time (BPP).

3 Progress in Optimization over networks

We have published the following results on optimization over social networks with Sebastien Roch.

On the Submodularity of Influence in Social Networks [12]. Social networks are often represented as directed graphs where the nodes are individuals and the edges indicate a form of social relationship. A simple way to model the diffusion of ideas, innovative behavior, or "word-of-mouth" effects on such a graph is to consider an increasing process of "infected" (or active) nodes: each node becomes infected once an activation function of the set of its infected neighbors crosses a certain threshold value. Such a model was introduced by Kempe, Kleinberg, and Tardos (KKT) in [8, 9] where the authors also impose several natural assumptions: the threshold values are random and the activation functions are monotone and submodular. The monotonicity condition indicates that a node is more likely to become active if more of its neighbors are active, while the submodularity condition indicates that the marginal effect of each neighbor is decreasing when the set of active neighbors increases.

For an initial set of active nodes S, let $\sigma(S)$ denote the expected number of active nodes at termination. Here we prove a conjecture of KKT: we show that the function $\sigma(S)$ is submodular under the assumptions above. We prove the same result for the expected value of any monotone, submodular function of the set of active nodes at termination. Roughly, our results demonstrate that "local" submodularity is preserved "globally" under this diffusion process. This is of natural computational interest, as many optimization problems have good approximation algorithms for submodular functions.

Another aspects of aggregating information from different sources in social networks is discussed in a recent work with G. Schoenebeck:

Reaching Consensus on Social Networks [14] In this work we study consensus algorithms on social networks. Research in sociology studies the effectiveness of social networks in achieving computational tasks. Typically the agents who are supposed to achieve a task are unaware of the underlying social network except their immediate friends. They have limited memory, communication, and coordination. These limitations result in computational obstacles in achieving otherwise trivial computational problems. One of the simplest problems studied in the social sciences involves reaching a consensus among players between two alternatives which are otherwise indistinguishable. In our paper we formalize the computational model of social networks. We then analyze the consensus problem as well as the problem of reaching a consensus which is identical to the majority of the original signals. In both models we seek to minimize the time it takes players to reach a consensus.

A different kind of problem involving optimization over social network is studied in a joint work with Costis Daskalakis and Alex Dimakis, two students from U.C. Berkeley. In this project we study the effect of the connectivity of a network in the context of random graphical games, where the main question is the existence of mutually optimal strategies, i.e., Nash Equilibrium.

Connectivity and Equilibrium in Random Games [7]. We study how the structure of the interaction graph affects the Nash equilibria of the resulting game. In particular, for a fixed interaction graph, we are interested if there exist Nash equilibria which arise when random utility tables are assigned to the players. We provide conditions for the structure of the graph under which equilibria are likely to exist and complementary conditions which make the existence of equilibria highly unlikely. Our results have immediate implications for many deterministic graphs and generalize known results for games on the complete graph. In particular, our results imply that the probability that bounded degree graphs have Nash equilibria is exponentially small in the size of the graph and yield a simple algorithm that finds small non-existence certificates for a large family of graphs. In order to obtained a refined characterization of the degree of connectivity associated with the existence of equilibria, we study the model in the random graph setting. In particular, we look at the case where the interaction graph is drawn from the Erdős-Rényi, $G(n,p)$, where each edge is present independently with probability p. For this model we establish a *double phase transition* for the existence of pure Nash equilibria as a function of the average degree pn consistent with the non-monotone behavior of the model. We show that when the average degree satisfies $np > (2+\Omega(1))\log n$, the number of pure Nash equilibria follows a Poisson distribution with parameter 1. When $1/n << np < (0.5-\Omega(1))\log n$ pure Nash equilibria fail to exist with high probability. Finally, when $np << 1/n$ a pure Nash equilibrium exists with high probability.

As part of our study we made an important progress in understanding how non-linearity coupled with sparsity can lead to optimal compression.

Smooth compression, Gallager bound and Nonlinear sparse-graph codes [11] A data compression scheme is defined to be *smooth* if its image (the codeword) depends gracefully on the source (the data). Smoothness is a desirable property in many practical contexts, and widely used source coding schemes lack of it.

We introduce a family of smooth source codes based on sparse graph constructions, and prove them to achieve the (information theoretic) optimal compression rate for a dense set of iid sources. As a byproduct, we show how Gallager bound on sparsity can be overcome

using non-linear function nodes.

4 Progress in Message Passing and MCMC algorithms

In a joint work with my student Allan Sly we considered the convergence rate of MCMC Gibbs samplers for sampling graph distributions for graphs that are sparse on average. Our results are the first to establish rapid convergence for random graphs. The results are in two papers, the first of which was accepted as an extended abstract to SODA 2008, while the second one has been submitted.

Rapid Mixing of Gibbs Sampling on Graphs that are Sparse on Average [15, 16]. Gibbs sampling also known as Glauber dynamics is a popular technique for sampling high dimensional distributions defined on graphs. Of special interest is the behavior of Gibbs sampling on the Erdős-Rényi random graph $G(n, d/n)$, where each edge is chosen independently with probability d/n and d is fixed. While the average degree in $G(n, d/n)$ is $d(1 - o(1))$, it contains many nodes of degree of order $\log n / \log \log n$.

The existence of nodes of almost logarithmic degrees implies that for many natural distributions defined on $G(n, p)$ such as uniform coloring (with a constant number of colors) or the Ising model at any fixed inverse temperature β, the mixing time of Gibbs sampling is at least $n^{1+\Omega(1/\log\log n)}$. Recall that the Ising model with inverse temperature β defined on a graph $G = (V, E)$ is the distribution over $\{\pm\}^V$ given by $P(\sigma) = \frac{1}{Z} \exp(\beta \sum_{(v,u)\in E} \sigma(v)\sigma(u))$. High degree nodes pose a technical challenge in proving polynomial time mixing of the dynamics for many models including the Ising model and coloring. Almost all known sufficient conditions in terms of β or number of colors needed for rapid mixing of Gibbs samplers are stated in terms of the maximum degree of the underlying graph.

In this work we show that for every $d < \infty$ and the Ising model defined on $G(n, d/n)$, there exists a $\beta_d > 0$, such that for all $\beta < \beta_d$ with probability going to 1 as $n \to \infty$, the mixing time of the dynamics on $G(n, d/n)$ is polynomial in n. Our results are the first polynomial time mixing results proven for a natural model on $G(n, d/n)$ for $d > 1$ where the parameters of the model do not depend on n. They also provide a rare example where one can prove a polynomial time mixing of Gibbs sampler in a situation where the actual mixing time is slower than $npolylog(n)$. Our proof exploits in novel ways the local treelike structure of Erdős-Rényi random graphs, comparison and block dynamics arguments and a recent result of Weitz.

Our results extend to much more general families of graphs which are sparse in some average sense and to much more general interactions. In particular, they apply to any graph for which every vertex v of the graph has a neighborhood $N(v)$ of radius $O(\log n)$ in which the induced sub-graph is a tree union at most $O(\log n)$ edges and where for each simple path in $N(v)$ the sum of the vertex degrees along the path is $O(\log n)$. Moreover, our result apply also in the case of arbitrary external fields and provide the first FPRAS for sampling the Ising distribution in this case. We finally present a non Markov Chain algorithm for sampling the distribution which is effective for a wider range of parameters. In particular, for $G(n, d/n)$ it applies for all external fields and $\beta < \beta_d$, where $d \tanh(\beta_d) = 1$ is the critical point for decay of correlation for the Ising model on $G(n, d/n)$.

Gibbs Rapidly Samples Colorings of $G(n, d/n)$ [18]. Gibbs sampling also known as Glauber dynamics is a popular technique for sampling high dimensional distributions

defined on graphs. Of special interest is the behavior of Gibbs sampling on the Erdős-Rényi random graph $G(n, d/n)$, where each edge is chosen independently with probability d/n and d is fixed. While the average degree in $G(n, d/n)$ is $d(1 - o(1))$, it contains many nodes of degree of order $\log n / \log \log n$.

The existence of nodes of almost logarithmic degrees implies that for many natural distributions defined on $G(n, p)$ such as uniform coloring (with a constant number of colors) or the Ising model at any fixed inverse temperature β, the mixing time of Gibbs sampling is at least $n^{1+\Omega(1/\log \log n)}$. High degree nodes pose a technical challenge in proving polynomial time mixing of the dynamics for many models including coloring. Almost all known sufficient conditions in terms of number of colors needed for rapid mixing of Gibbs samplers are stated in terms of the maximum degree of the underlying graph.

In this work consider sampling q-colorings and show that for every $d < \infty$ there exists $q(d) < \infty$ such that for all $q \geq q(d)$ the mixing time of Gibbs sampling on $G(n, d/n)$ is polynomial in n with high probability. Our results are the first polynomial time mixing results proven for the coloring model on $G(n, d/n)$ for $d > 1$ where the number of colors does not depend on n. They also provide a rare example where one can prove a polynomial time mixing of Gibbs sampler in a situation where the actual mixing time is slower than $npolylog(n)$. In previous work we have shown that similar results hold for the ferromagnetic Ising model. However, the proof for the Ising model crucially relied on monotonicity arguments and the "Weitz tree" both of which have no counterparts in the coloring setting. Our proof presented here exploits in novel ways the local treelike structure of Erdős-Rényi random graphs, block dynamics, spatial decay properties and coupling arguments.

Our results give first FPRAS to sample coloring on $G(n, d/n)$ with a constant number of colors. They extend to much more general families of graphs which are sparse in some average sense and to much more general interactions. In particular, they apply to any graph for which there exists an $\alpha > 0$ such that every vertex v of the graph has a neighborhood $N(v)$ of radius $O(\log n)$ in which the induced sub-graph is a tree union at most $O(1)$ edges and where each simple path Γ of length $O(\log n)$ satisfies $\sum_{u \in \Gamma} \sum_{v \neq u} \alpha^{d(u,v)} = O(\log n)$. The results also generalize to the hard-core model and other models where there are both hard and soft constraints.

The two results above are complemented by the following breakthrough result where the *exact* threshold for mixing of Gibbs samplers for Ising model is found.

Exact Thresholds for Ising-Gibbs Samplers on General Graphs [17] We establish tight results for rapid mixing of Gibbs Samplers for the Ferromagnetic Ising model on general graphs. We show that if

$$(d - 1) \tanh \beta < 1,$$

then there exists a constant C such that the discrete time mixing time of Gibbs Samplers for the Ferromagnetic Ising model on *any* graph of n vertices and maximal degree d, where all interactions are bounded by β, and arbitrary external fields is bounded by $Cn \log n$. Moreover, the spectral gap is uniformly bounded away from 0 for all such graphs as well as for infinite graphs of maximal degree d.

We further show the when $d \tanh \beta < 1$, with high probability over the Erdős-Rényi random graph $G(n, d/n)$, it holds that the mixing time of Gibbs Samplers is

$$n^{1+\Theta(\frac{1}{\log \log n})}.$$

Both result are tight as it is known that the mixing time for random regular and Erdős-Rényi random graphs is, with high probability, exponential in n when $(d-1)\tanh\beta > 1$, and $d\tanh\beta > 1$, respectively. To our knowledge our results give the first tight sufficient conditions for rapid mixing of spin systems on general graphs. Moreover, our results are the first rigorous results establishing exact thresholds for dynamics on random graphs in terms of spatial thresholds on trees.

We also study message passing algorithm: In a joint work with Coja-Oghlan and Vilenchik we show how spectral techniques explain the success of Belief Propagation for solving coloring problems.

A Spectral Approach to Analyzing Belief Propagation for 3-Coloring [6]. Contributing to the rigorous understanding of BP, in this paper we relate the convergence of BP to spectral properties of the graph. This encompasses a result for random graphs with a "planted" solution; thus, we obtain the first rigorous result on BP for graph coloring in the case of a complex graphical structure (as opposed to trees). In particular, the analysis shows how Belief Propagation breaks the symmetry between the 3! possible permutations of the color classes.

The next problem we study is the power of greedy algorithms in inferring random variables on networks.

Iterative Maximum Likelihood on Networks [19] We consider n agents located on the vertices of a connected graph. Each agent v receives a signal $X_v(0) \sim N(\mu, 1)$ where μ is an unknown quantity. A natural iterative way of estimating μ is to perform the following procedure. At iteration $t+1$ let $X_v(t+1)$ be the average of $X_v(t)$ and of $X_w(t)$ among all the neighbors w of v. It is well known that this procedure converges to $X(\infty) = \frac{1}{2}|E|^{-1}\sum d_v X_v$ where d_v is the degree of v.

In this paper we consider a variant of simple iterative averaging, which models "greedy" behavior of the agents. At iteration t, each agent v declares the value of its estimator $X_v(t)$ to all of its neighbors. Then, it updates $X_v(t+1)$ by taking the maximum likelihood (or minimum variance) estimator of μ, given $X_v(t)$ and $X_w(t)$ for all neighbors w of v, and the structure of the graph.

We give an explicit efficient procedure for calculating $X_v(t)$, study the convergence of the process as $t \to \infty$ and show that if the limit exists then $X_v(\infty) = X_w(\infty)$ for all v and w. For graphs that are symmetric under actions of transitive groups, we show that the process is efficient. Finally, we show that the greedy process is in some cases more efficient than simple averaging, while in other cases the converse is true, so that, in this model, "greed" of the individual agents may or may not have an adverse affect on the outcome.

The model discussed here may be viewed as the Maximum-Likelihood version of models studied in Bayesian Economics. The ML variant is more accessible and allows in particular to show the significance of symmetry in the efficiency of estimators using networks of agents.

4.1 Progress on Learning from Multiple Sources

We finally briefly note work with J. Arpe on learning from multiple sources:

Application of a Generalization of Russo's Formula to Learning from Multiple Random Oracles [1] In this paper we study the problem of learning low complexity

functions called *k-juntas* given access to examples drawn from a number of different product distributions. Thus we wish to learn a function $f : \{-1,1\}^n \to \{-1,1\}$ that depends on k (unknown) coordinates. While the best known algorithms for the general problem of learning a k-junta require running time which exceeds n^k, we show that given access to k different product distributions with biases separated by $\epsilon > 0$, the functions may be learned in time which is polynomial in n and k. Our techniques involve novel results in Fourier analysis relating Fourier expansions with respect to different biases and a generalization of Russo's formula.

References

[1] J. Arpe and E. Mossel. Application of a generalization of russo's formula to learning from multiple random oracles. *Combinatorics, Probability and Computing*, 19(2):183–199, 2010.

[2] A. Bogdanov, E. Mossel, and S. Vadhan. The complexity of distinguishing markov random fields. In *11th International Workshop, APPROX 2008, and 12th International Workshop, RANDOM 2008, LNCS 5171*, pages 331–342. Springer, 2008.

[3] M. Braverman and E. Mossel. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 268–276, 2008.

[4] M. Braverman and E. Mossel. Sorting from noisy information. Arxiv 0910.1191, 2010.

[5] G. Bresler, E. Mossel, and A. Sly. Reconstruction of markov random fields from samples: Some easy observations and algorithms. In *11th International Workshop, APPROX 2008, and 12th International Workshop, RANDOM 2008, LNCS 5171*, pages 343–356. Springer, 2008.

[6] A. Coja-Oghlan, E. Mossel, and D. Vilenchik. A spectral approach to analyzing belief propagation for 3-coloring. *Combinatorics, Probability and Computing*, 18(6):881–912, 2009.

[7] C. Daskalakis, A. G. Dimakis, and E. Mossel. Connectivity and equilibrium in random games. Arxiv 0703.5902, 2010.

[8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2003.

[9] D. Kempe, J. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *Proc. 32nd International Colloquium on Automata, Languages and Programming (ICALP)*, 2005.

[10] F. A. Matsen, E. Mossel, and M. Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bull. Math. Bio.*, 70(4):1115–1139, 2008.

[11] A. Montanari and E. Mossel. Smooth compression, gallager bound and nonlinear sparse graph codes. 2008. In Proceedings of ISIT 2008.

[12] E. Mossel and S. Roch. On the submodularity of influence in social networks. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 128–134, 2007.

[13] E. Mossel and S. Roch. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *IEEE Comp. Bio. and Bioinformatics*, 7(1):166–171, 2010.

[14] E. Mossel and G. Schoenebeck. Reaching consensus on social networks. In *Proceedings of 1st Symposium on Innovations in Computer Science*, pages 214–229, 2010.

[15] E. Mossel and A. Sly. Rapid mixing of gibbs sampling on graphs that are sparse on average. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 238–247, 2008.

[16] E. Mossel and A. Sly. Rapid mixing of gibbs sampling on graphs that are sparse on average. *Random Structures and Algorithms*, 35(2):250–270, 2009.

[17] E. Mossel and A. Sly. Exact thresholds for ising-gibbs samplers on general graphs. Arxiv 0903.2906, 2010.

[18] E. Mossel and A. Sly. Gibbs rapidly samples colorings of g(n,d/n). *PTRF*, 48:37–69, 2010.

[19] E. Mossel and O. Tamuz. Iterative maximum likelihood on networks. In *Proceedings of Forty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, 2009.